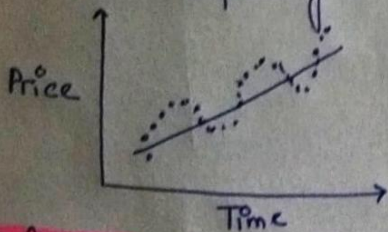
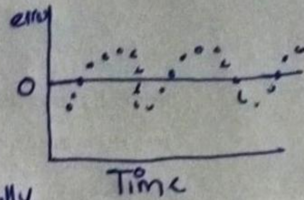
# AUTO-CORRELATION / SERIAL CORRELATION

Consider the following stock chart for a company.




$$\text{Stock price} = \beta_0 + \beta_1(\text{Time}) + \varepsilon \xrightarrow[\text{plot}]{\text{Residual}}$$




Residual plot the best fit line horizontally at 0.

In this we can see a snake pattern that is auto correlation. (Repeating same pattern over period of time)

**Definition** → Autocorrelation also known as serial correlation is the correlation of a signal with delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of time lag between them. The analysis of autocorrelation is a mathematical tool for finding repeating patterns such as periodic signal obscured by noise or identifying the missing fundamental frequency in a signal.

→ Autocorrelation represent the degree of similarity between a given time series and lagged version of itself over successive time intervals.

→ Auto correlation measures the relationship between a variable's current value and its past values.

→ An auto correlation of +1 represent a perfect positive correlation, while an autocorrelation of −1 represent a perfect negative correlation.

→ For example, auto correlation can help us to see how much of an impact past prices for a security have on its future price. Autocorrelation can show if there is momentum factor associated with stock. For eg, if investors knows that a stock that has a historically high positive autocorrelation value and they witness it making sizeable gains over past several days, then they might reasonably expect the movements over the upcoming several days.

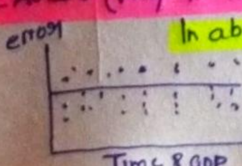
→ Another example, one might expect the air temperature on the 1st day of the month to be more similar to the temperature on 2nd day compared to 31st day. If the temperature values that occured closer together in time are, in fact more similar than the temperature values that occured farther apart in time, the data will be auto correlated.

**PROBLEM** → In regression analysis, autocorrelation of the regression residuals can also occur if the model is incorrectly specified. If For example, if we are attempting to model a simple linear relationship but the observed relationship is non-linear then residuals can be auto correlated.

**CAUSES (Why it cause)** — Cause 1: Ommitted Variables

In above example of $\text{stock price} = \beta_0 + \beta_1(\text{Time}) + \varepsilon_i$, suppose only time not able to derive stock price correctly but if we add GDP then it is able to predict correctly.

$$\text{stock price} = \beta_0 + \beta_1(\text{Time}) + \beta_2(\text{GDP}) + \varepsilon_i$$

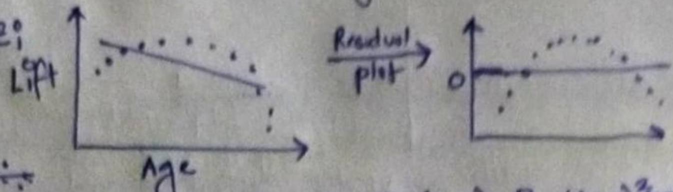



→ By adding GDP, now error are normally distributed in residual plot so either we should check remove variable or we should add new variable.

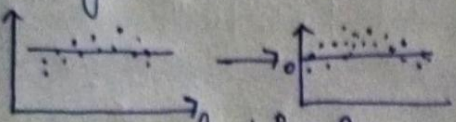
## CAUSE II → Incorrect functional form.

Suppose, how much ~~person~~ weight a person can lift (For weightlifters)

Max weight lift = $\beta_0 + \beta_1 (Age) + \varepsilon_i$




Suppose if we add $(Age)^2$, then it can resolve the problem.

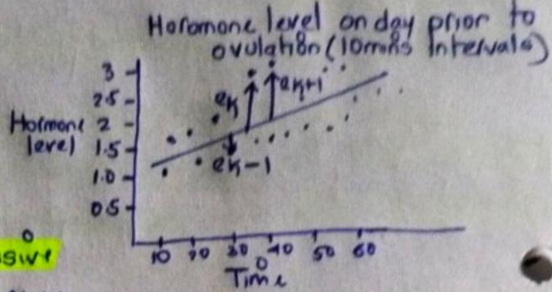



→ Max weight = $\beta_0 + \beta_1 (Age) + \beta_2 (Age)^2 + \varepsilon_i$

So, simply we corrected the functional form.

## DIAGNOSIS AUTO CORRELATION

### Diagnosis 1 – Durbin Watson Test




→ Created in 1950.

→ Run the reregression and caputure error terms.

→ So we check successive error terms if they are related. For eg, $e_k$ and $e_{k+1}$

→ Positive auto correlation we can say if two successive point have positive error. For example, $e_k$ and $e_{k+1}$

→ Negative autocorrelation we say if two successive points have opposite error. For eg, $e_{k-1}$ & $e_k$

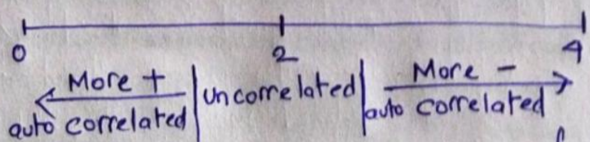
→ Durbin Watson test can be applied only on successive term (Only first order autocorrelation)
It cannot go with second order ($e_k$, $e_{k+2}$) or third order autocorrelation ($e_k$, $e_{k+3}$)

→ Calculate dublin watson statistics, $dw = \dfrac{(e_2-e_1)^2 + (e_3-e_2)^2 + \dots + (e_n - e_{n-1})^2}{e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2}$

$$dw = \frac{\left(\sum_{i=2}^{n} (e_i - e_{i-1})^2\right)}{\sum_{i=1}^{n} e_i^2}$$

→ Scale of dublin watson statistic → 0 to 4.

Cut offs depend on:
i) n : number of observation
ii) k : number of x variables




Suppose, if we apply DW test on above problem and DW = 1.312. Therefore we can conclude positive autocorrelation exists.
So it means if there is positive error term then the next term will be also positive error term ie, $e_k$ and $e_{k+1}$. And if there is a negative error term then it will be also followed by negative error term.
It means we cannot rely on regression output especially t value in the regression output (due to variance)

→ Note – It only check first order autocorrelation.

## DIAGNOSIS 2 – Breusch – Godfrey test (BG test)

→ Created in 1978

Hormone level$_t$ $= \beta_0 + \beta_1(\text{Time}) + \varepsilon_t$.

- Get error terms from original regression $e_t$ (sample error) / $\varepsilon_t$ (population error)
- So BG test tells us to run auxilary regression, find error term ($e_t$) & run regression

$$e_t = \beta_0 + \beta_1(\text{time}) + \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \gamma_3 e_{t-3} + \dots + \gamma_p e_{t-p} + \mu_t$$

based on previous error term like $t-1$, $t-2$, $t-3$ ... etc.

There will be $\mu_t$ (last term) which will not be affected by auto correlation.

→ If Time variable affects the error term, then it is called endogenous. Endogenous is simply a variable which is omitted that is affecting time variable as well as error term.

→ So by run auxilary regression, we trying to find effect of omitted variable.

→ So we can asses by the R-squared of auxilary regression. If R square is high then current error term is related to previous error term, so it will suffer from autocorrelation.

Normally, null hypothesis – No autocorrelation.

Alternate hypothesis – There is a autocorrelation.

## REMEDIES OF AUTOCORRELATION

① Add in ommitted variable (If any) ② Correct any functional form issues ③ Create a general difference equation

③ Create a generalised Difference Equation.

Hormone level$_t$ (HL) $= \beta_0 + \beta_1(\text{time}_t) + \varepsilon_t$ ⟶ Error term ($\varepsilon_i$) is function of one before it, that what autocorrelation implies.

$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t.$$

$\rho$ tells how error ($t-1$) is autocorrelated with $\varepsilon_t$

$\mu_t$ is ~~unrrelate~~ unrelated error term.

$$\mu_t = \varepsilon_t - \rho\varepsilon_{t-1}.$$

So we have to find $\mu_t$ (Main aim)

$$HL_t = \beta_0 + \beta_1(\text{time}_t) + \varepsilon_t \quad \text{—(i)}$$

For $t-1$

$$HL_{t-1} = \beta_0 + \beta_1(\text{time}_{t-1}) + \varepsilon_{t-1} \quad \text{—(ii)}$$

Multiply $\rho$ to (ii)

$$\rho(HL_{t-1}) = \rho\beta_0 + \rho\beta_1(\text{Time}_{t-1}) + \rho\varepsilon_{t-1} \quad \text{—(iii)}$$

eqn (i) – (iii)

$$HL_t - \rho(HL_{t-1}) = \beta_0 + \beta_1(\text{time}_t) + \varepsilon_t - \rho\beta_0 - \rho\beta_1(\text{Time}_{t-1}) - \rho\varepsilon_{t-1}$$

$$= \beta_0(1-\rho_0) + \beta_1\text{time}_t - \rho\beta_1(\text{time}_{t-1}) + \underline{\varepsilon_t - \rho\varepsilon_{t-1}}$$

For $\rho$ use AR(1) everything will be settled.